Do we really need ‘Big Data” for patient-level predictive modelling?

Peter R. Rijnbeek, PhD[1,3], Xiaoyong Pan, PhD[1,3], Jenna Reps, PhD[2,3]
[1]Erasmus University Medical Center, Rotterdam, Netherlands, [2]Janssen Research and Development, Raritan, NJ
[3] Observational Health Data Sciences and Informatics (OHDSI), New York, NY, USA

Background

The standardization to the OMOP-CDM and the development of the patient-level prediction framework has opened up unprecedented possibilities for large-scale patient-level predictive modelling using Electronic Health Records (EHRs). We can now run large-scale analyses to explore the predictive performance across numerous algorithms, databases, target populations and outcome combinations. However, training models using large datasets is time consuming.Therefore, we have added learning curve functionality to the PatientLevelPrediction package to assess the effect of training size.

A learning curve has traditionally been used to assess if collecting more costly data would help to improve model performance. It provides insight in whether the model suffers from high bias or high variance [1,2]. If the performance of the train and test set converges to a low performance the model suffers from high bias and more data will not improve model performance. The only possibility to obtain a potentially better performing model is to increase its complexity. However, in the case of a model suffering from high variance, more data often helps. Interestingly, in OHDSI we have a slightly different situation: we have access to very large datasets at a low cost, but the question is:

Do we really need to run our models on millions of patients for many hours?

Our hypothesis is that this is not the case but we need to understand better how to determine the optimal training set size to balance computational load and model performance. These insights would make our all-by-all objective, i.e. develop predictive models for all target cohorts and all outcomes, possibly more realistic.

Methods

To test our hypothesis, we have added learning curve functionality to the PatientLevelPrediction package which allows us to run every available model in the package on increasing training set size using the same test set. The full dataset only needs to be extracted once and the function creates stratified subsamples of the training set of increasing size based on the user defined list of fractions. A plotting function is added to visualize the training and test performance. As a proof-of-concept we developed learning curves on a very large (>120 million people) claims database (CCAE) for four outcomes in a cohort of patients with pharmaceutically treated depression. We developed **Lasso and Random Forest models using 1,2,4,8,10,20,30,40,50,60,70,80% of the persons as training set and used the same set of 20% of the persons as test set**. For both models, the default hyper parameter searches were used. The outcomes of interest are: AMI, Hypothyroidism, Suicide, and Stroke.

Results

In Figure 1, the results are shown for the four outcomes. In each panel, the Lasso and Random Forest performance on the train and test set are shown as function of the number of subjects with the outcome in the training sample. These curves provide some interesting insights:

1. For the two large outcome cohorts (Hypothyroidism, and Suicide) we could have stopped much earlier.

2. For Hypothyroidism, Random Forest seems to stabilize earlier than lasso.

3. In the case of AMI and Stroke more data would still help since we are still in the increasing part of the learning curve.

4. For all these outcomes Lasso is working best and the models are far less overfitted. This may be because the default grid-search is suboptimal for RF and small depth values still need to be added.

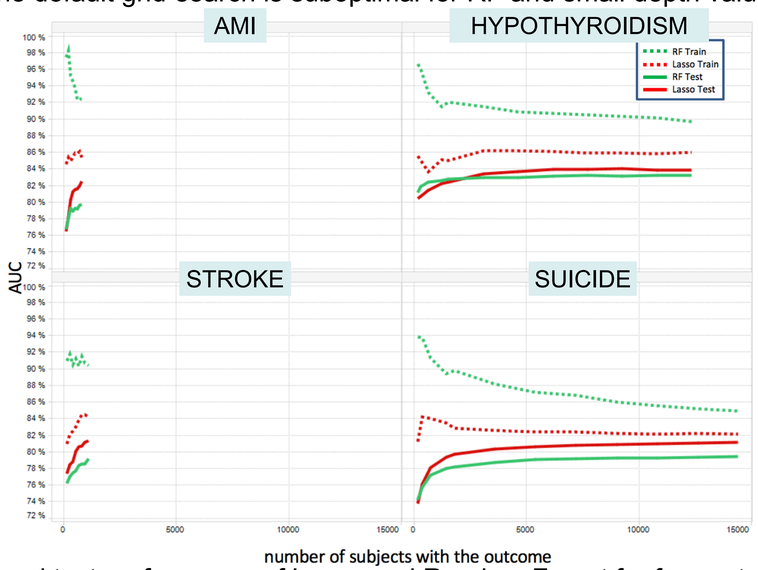

Figure 1. Train and test performance of Lasso and Random Forest for four outcomes as function of the number of subjects with the outcome in the training set.

Conclusions

Learning curves are helpful for assessing the required training set size for patient-level prediction. We could develop a procedure that only enlarges the training sample if needed. Ideally, we would like to be able to make an educated guess based on more learning curve experience. We therefore want to extend the current study to many cohorts at risk and outcomes in multiple large databases.

References
1. D. Brain and G. Webb. On the effect of data set size on bias and variance in classification learning. In Proceedings of the Fourth Australian Knowledge Acquisition Workshop, University of New South Wales, pages 117–128, 1999.
2. Perlich, C., Provost, F. and Simonoff, J., (2003), “Tree Induction vs. Logistic Regression: A Learning-curve Analysis”, Journal of Machine Learning Research, 4, 211–255.

Contact:p.rijnbeek@erasmusmc.nl